**HW2 Update on Analysis of Water Consumption and Cost in New York City**

**Hypothesis**

Higher water consumption increases water costs, but cost per unit of water varies depending on billing periods, location, and potential penalties for high usage.

**Problem Statement**

This project analyzes the relationship between water usage and cost to determine how different factors influence pricing. The goal is to discover whether cost per unit is consistent or affected by external elements such as location, billing cycles, or minimum service charges.

**Progress Overview**

**1. Repository Management**

* GitHub Repository: [Water\_Consumption\_Analysis](https://github.com/wcalderon1/Water_Consumption_Analysis)
* Cloned to local system using:

bash

git clone https://github.com/wcalderon1/Water\_Consumption\_Analysis.git

**2. Data Preparation**

* Dataset source: NYC Open Data portal
* Downloaded using:

bash

curl 'https://data.cityofnewyork.us/api/views/66be-66yr/rows.csv?accessType=DOWNLOAD' -o nyc\_water\_data.csv

* Extracted relevant columns using:

bash

awk -F, '{print $24, $25}' nyc\_water\_data.csv > consumption\_costs.csv

* Saved cleaned version as: cleaned\_consumption\_costs.csv

**3. Hadoop Processing**

* Uploaded data to HDFS:

bash

hdfs dfs -put /home/data/cleaned\_consumption\_costs.csv /user/data/

* Ran MapReduce job with custom driver:

bash

hadoop jar /home/data/WaterConsumptionAnalysis.jar WaterAnalysisDriver /user/data/cleaned\_consumption\_costs.csv /user/data/output

* Verified output:

bash

hdfs dfs -cat /user/data/output/part-00000 | head -n 10

**Updated Conclusions**

* **Positive Correlation**: The results show a strong positive relationship between water consumption and total charges.
* **Cost per Unit Varies**: Different rows with the same consumption had different charges, suggesting the presence of tiered pricing or other billing policies.
* **Anomalies Found**: Several entries had zero consumption but non-zero charges, possibly due to base fees or penalties.
* **Distribution Patterns**: Some charges cluster around specific price points, supporting the idea of billing tiers.

**Hadoop Processing & Final Conclusions**

The cleaned dataset (cleaned\_consumption\_costs.csv) was uploaded to HDFS and processed using a custom Hadoop MapReduce job. The job was compiled into a .jar file (WaterConsumptionAnalysis.jar) and executed using:

bash

hadoop jar WaterConsumptionAnalysis.jar WaterAnalysisDriver /user/data/cleaned\_consumption\_costs.csv /user/data/output

The job completed successfully and generated an output file (part-00000) in HDFS, which was retrieved using:

bash

hdfs dfs -get /user/data/output/part-00000

This file was added to the GitHub repository and committed as part of the project:

bash

git add part-00000

git commit -m "Add Hadoop part-00000 output"

git push origin main

**Final Conclusions**

* The MapReduce job successfully grouped water consumption values by their corresponding water and sewer charges.
* The result shows that similar charges can be associated with a wide range of consumption levels, confirming that cost per unit is not consistent.
* Charges with zero consumption still appeared, reinforcing the idea of flat fees or base charges in the billing structure.
* The hypothesis is **partially supported**: while higher consumption tends to raise costs, billing rules and potential fees introduce variability.

**GitHub Repository:**  
 <https://github.com/wcalderon1/Water_Consumption_Analysis>